MSS SHORT USER'S GUIDE

Version 1.0

February 5, 1998

Ken Fidler, Krzysztof Genser, Jeff Mack, Alexander Moibenko, David Sachs, Joseph Syu

Mass Storage Systems Group

High Performance and Parallel Computing Department

Computing Division

Fermilab

Table of contents

1. What is MSS?

The Mass Storage System (MSS) is a central facility that was developed and maintained by the MSS group, which is part of the HPPC department of the Computing Division. This facility was developed to provide standardized ways of storing and retrieving large bulk data. Various tools are available to users to access and organize their data within the Mass Storage System.

The MSS is comprised of several large components. One major component is the Hierarchical Storage Management (HSM) system. The HSM used by the MSS facility appear as a large UNIX based file system. In reality, the HSM is more than just a file system. HSM allow us to combine various data storage technologies to provide an organized data repository. By combing high-speed networks, high-speed disk, and robotics tape storage facilities, an HSM can provide easy access to multi terabytes of data. To utilize an HSM for the needs of the Fermilab Scientific community, additional components are combined with the HSM to form the MSS.

The HSM provides the necessary organization and control of the data across a scaleable system that incorporates the robotics tape library. The MSS also includes a sub-system to import and export user data to and from the HSM. In addition various tools were built that aid the user community in organizing and accessing data to and from the HSM. The MSS also includes additional fault tolerance and security measures that are layered on top of the HSM.

2. Quick Start Procedures

Please substitute parameters in Italics with the appropriate information.

2.1 Putting 8mm tape data into the MSS

Issue the following interactively from FNALU IBM or IRIX systems, or submit a batch job to the CLUBS batch facility with one of the following:

```
needfile -stage -vsn xxxxxx -tape 8mm (single density/unlabeled)
needfile -stage -vsn xxxxxx -tape 8mma (single density labeled)
needfile -stage -vsn xxxxxx -tape 8mmd (double density unlabeled)
needfile -stage -vsn xxxxxx -tape 8mmda (double density labeled)
```

where *xxxxxx* is the name of the tape.

2.2 Putting disk files into the MSS

Issue the following from UNIX system either in batch or interactively:

```
fmss cp disk:my_file mss:group/path/new_file_name
```

2.3 Retrieving data out of the MSS onto local disk

Issue the following from UNIX system either in batch or interactively:

For data that was stored via needfile, issue:

```
fmss cp needfile:vsn/file_number disk:output_name
```

For data that was stored via fmss, issue:

```
fmss cp mss:group/path/file_name disk:output_name
```

2.4 Deleting data from the MSS

Issue the following from UNIX system either in batch or interactively:

```
fmss rm mss:group/path/file-name
```

Please note that the fmss commands can not neither be reverted nor stopped.

3. Getting access to MSS

Access to the MSS is made available to those user groups that are approved by the Computing Division. It is not necessary for each member of an experiment group to have an account on the MSS. One account must be defined as a controlling account for the MSS system for that particular group. This account is used to control and establish ownership of a particular directory structure. The MSS will appear as one large file system. Each group that is granted permission to the MSS will have a directory path defined. For example, experiment E999 could have a directory path called E999. With the tools provided, users can then add standard UNIX directory paths under the controlling main path. For example, E999 could define a sub-directory called RUN123 under E999. The use of UNIX directory paths provides a flexible organization structure that can be defined by the experiment group. The tools developed for the MSS allow access of experiments' data by any member of the experiment group as well as non-members of the group. In addition, the tools provide access to the MSS without forcing users to have to know any passwords.

The design of the MSS also provides experiment groups the ability to define additional accounts for individual users of the experiment group. This allows a particular member of a group to have a 'private' area within the experiment group's area. For example, a particular member of the group may need to restrict access to a particular group of data. Or, a group may wish to establish an area maintained by a small group of people with the rest of the group having just read access to the files. By providing a special account area, the user can define specific UNIX like file permissions on his/her data. This method can also be used for sub-groups within an experiment. Additional accounts for a group area must be created by the MSS Group.

The security feature within the MSS is based upon the standard UNIX user and group ids that are defined by the Computing Division. On trusted systems such as FNALU, if a user is defined to be a member of a particular group, the user will then have access to the defined MSS area for your group. On non-trusted systems, since the User and Group ids may not be the same as defined by the Computing Division, a facility similar to the UNIX '.rhosts' file concept was provided. A special file called the '.access' can be created by members of the experiment group to add additional users and the specific nodes these users can come from. Users defined in this file are granted the same capabilities as members of the group from the trusted hosts.

4. Tools to access MSS

The following is a brief summary of the tools available to inter-operate with the MSS. Following the summary, is additional information on when to use the various tools for your various needs:

4.1 fmss

fmss is the main tool that to be used to get and put data into the MSS. This tool is available via Fermilab UPS/UPD and is designed to work on IBM AIX, SGI IRIX, SUN OS, Digital UNIX and Linux operating systems. The product can be 'setup' from any of the FNALU and CLUBS nodes by issuing 'setup fmss'. For more details on fmss, please review fmss documentation available on the HPPC department home WWW page (http://www-hppc.fnal.gov/mss).

Note: fmss is a continually being enhanced to incorporate additional functionality. Some of the following tools will be replaced with future enhancements in the fmss product.

4.2 needfile

needfile is a product used to import user data from 8mm tapes into the MSS using a robust central facility. Data imported in by needfile is maintained in a separate directory structure of the MSS, which is not owned by any experiment group. Data imported by needfile is maintained in the MSS with the least-recently-used algorithm, hence data that is imported into the MSS via needfile will automatically be removed if the data is not accessed over a period of time. The lengths of time can very depending upon system activity, but on an average, data remains in the system for about two weeks.

needfile is unique in that it provides the ability to import data into the MSS as well as retrieve the data. Currently fmss can retrieve data that was imported by the needfile system. needfile data is organized by tape VSN and each file on the tape is saved as an individual file. This product can be used to import 8mm tapes that are either standard labeled or non-labeled.

needfile is automatically setup for batch jobs on CLUBS. To use needfile on FNALU, one must be on either an SGI IRIX or and IBM AIX node. 'setup nt' enables use of needfile on FNALU system. The product nt is not available via Fermilab UPD, and should not be used on systems outside of CLUBS and FNALU. There are no plans to port needfile to any other platform.

fmss 'needfile: ' area provides an alternative way of accessing files imported via needfile.

4.3 mssout/clubsout/capout

mssout is a newly available product complementary to and to be included in the fmss that allows to copy files residing in the MSS to an 8mm tape. mssout as of now is available as a Fermilab UPS/UPD product for AIX, IRIX+5, SunOS+5, and OSF1 flavors of UNIX. mssout replaces the previous clubsout and capout products, and accepts (with minor changes) control files used by them. The data that is written to the output tape must have been stored in the MSS via fmss. mssout -q give the list of current mssout jobs. mssout builds a batch job that is submitted to the CLUBS batch system.

Clubsout and capout products were phased out by mssout.

4.4 tcache

tcache product was phased out by fmss 'mss:shorterm' area.

4.5 ndflist

ndflist is a tool that can be used to obtain information about data that is in the needfile portion of the MSS. Most of the functionality of ndflist is available in fmss ls. ndflist is setup by 'setup nt' on either the SGI IRIX or IBM AIX nodes on FNALU.

5. Putting data into MSS

5.1 Disk Data

For data, which already resides on disk, the best choice is to use the fmss command. fmss has a copy function called 'cp' which is very similar to the standard UNIX cp command. The major difference is that fmss deals with input and output areas that might not be visible from the system fmss is run on. fmss uses a concept called 'area type' to distinguish between different locations of data. Area type called 'mss:' is used to mean the Mass Storage System, 'disk:' area type denotes a data area that is directly accessible. For example, to copy a file from the current directory to the MSS system one would enter:

```
fmss cp disk:myfile mss:E999/files/myfile
```

This will copy the file called 'myfile' that is located in the current directory to the Mass Storage System in the E999 group area under the directory files.

Please refer to the fmss documentation for more specific details and options.

5.2 8mm Tape Data

If the data is on a labeled or unlabeled 8mm tape (either an 8200 or 8500 tape), one can use the Tape Import Subsystem of the MSS or read the tape locally on some other facilities. To use the MSS facility, one needs to use the 'needfile' command. One can either submit a batch job to the CLUBS system, or interactively issue the needfile command on either the IBM AIX or SGI IRIX nodes of the FNALU cluster. The basic form of the needfile command to cause the importation of 8mm tape is (single density/unlabeled is the default, so -tape may be an important parameter)

```
needfile -stage -vsn xyz345 -tape 8mm (single density/unlabeled)
needfile -stage -vsn xyz345 -tape 8mma (single density labeled)
needfile -stage -vsn xyz345 -tape 8mmd (double density unlabeled)
needfile -stage -vsn xyz345 -tape 8mmda (double density labeled)
```

In the example above, the system will import the tape vsn XYZ345 from the first file to and including all the other files until a double tape mark or the physical end of the tape. Each file will be stored in the MSS under the needfile directory, under the data format of stream, under the name of XYZ345. For an unlabeled tape, each file from the tape will be given the name DATAnnnn where nnnn is the file number of the file. For example, the 5th file on the tape would be called DATA0005. For labeled tapes, the system will use the name in the standard label for the name of the file.

6. Retrieving data out of MSS

The best way to retrieve data from the MSS is to use the fmss command. The fmss product can get data that was copied into the MSS using either fmss or needfile. The fmss command can be run from any node on the FNALU/CLUBS cluster. The fmss product can also be installed on any of the five UNIX platforms (AIX, SUN OS, IRIX, Digital UNIX and Linux). fmss is flexible

to allows to specify the location and the name of the retrieved files. It can be used in batch jobs on FNALU/CLUBS as well. For batch queues that provide a local work space, one should remember to specify the path. For example, to obtain a file from the MSS in a group area, one would use the following command:

```
fmss cp mss:E999/files/the_big_file disk:$FERMINT_DPOOL_DIR/my_file
```

In the case when data was imported via the needfile facility, one can use fmss to retrieve the data. For example, to obtain the third file from the tape XYZ123, one could issue the following:

```
fmss cp needfile:XYZ123/3 disk:my_directory
```

This would return the 3rd file from tape XYZ123 and place a copy of the file in the directory my_directory. The file would be named DATA0003 or the file would have the name that was supplied with the label if the tape was a labeled tape. With fmss cp, one can specify a different output file name. For example:

```
fmss cp needfile:XYZ123/3 disk:my_directory/somedata
```

This would do the same as the prior example, except the file would be renamed 'somedata'.

7. Removing data from MSS

Data that is stored into group area via the fmss command will remain on the system permanently. If one no longer needs the data, one can use the fmss rm command to delete the file. For example, one could issue the following:

```
fmss rm mss:E999/files/myfile
```

which would delete the file called myfile from the directory files in the E999 group area. Please note that the fmss commands can not neither be reverted nor stopped.

If the data was stored via the needfile process, data in the needfile area is automatically removed once it is not used for a certain amount of time. Data remains in the needfile area as long as the data is active, and there is room for the data. If one determines one no longer need the data, one can issue a needfile command to remove the data from the system:

```
needfile -free -vsn XYZ211
```

This will inform the system to remove the entire data set for the vsn XYZ211. One can not remove individual files from the VSN data set.

8. Write output tapes from data in MSS

A facility exists that allows to create and run a special batch job on the CLUBS tape server that will retrieve data from the MSS and write it to an 8mm tape. It is available as a Fermilab UPS product called `mssout`. It can be used to write out fmss and tcache data sets. It requires the user to create a special file that lists the MSS data sets and the vsn of the output tape. Mail will be sent back to the user when the mssout job is completed.

For example, one could create the following file called my_control_file:

```
mail=jack@fnal.gov
vsn=XYZ123
file=mss:E999/files/myfile.1
file=mss:E999/files/myfile.2
file=mss:E999/run123/filea
```

and execute the mssout command (after doing setup mssout) as follows:

```
mssout my_control_file
```

which will spool files myfile.1, myfile.2 and filea to the labeled tape XYZ123. If one wishes to have an ANSI labeled tape (which supports long VMS like file names) one should use an additional directive in the control file, namely:

```
labels=ANSI
```

mssout jobs can be listed using mssout -q command. The full description of mssout can be found at: http://fnhppc.fnal.gov/mss

9. Running out MSS space.

The HSM systems used are limited by the capacity available in the attached tape robotics libraries. To ensure that each group gets a fair allotment of the space available, a quota for each group is established. The fmss product keeps track of your data and can be used to determine how much of your quota your group has occupied. If the group attempts to exceed their quota, the fmss product will reject any new data until there is room available. fmss quota command can be used to determine group's quota.

For data that is imported via the needfile system, a maximum quota for the needfile system is monitored on a regular cycle. When the needfile system determines it is near its quota, the system automatically deletes the least recently used data sets. An entire set of files for a tape vsn is removed at a time.

10. What happens if MSS is down?

Both the fmss and needfile commands are based on the common client/server style. Every attempt has been made to include as much fault tolerance and recovery into the code. Occasionally network or other various conditions can cause recovery processing to occur. If errors occur, error information will be displayed back to the user. The user has the option of allowing the system to recover, or the user can issue a CTRL-C to abort the request. At any time, the user can issue a CTRL-C to cancel his/her request. Users can also issue the fmss status command to determine if the server is available. If the server is down for maintenance, a message will be sent back to the user indicating information regarding the maintenance outage.

fmss returns non zero return codes for irrecoverable errors. It is advisable to always check the value of the return code.

11. Examples

11.1 Looping through an 8mm tape loaded via the needfile command

```
#!/bin/ksh
#the next line helps setup “clean” environment
. /usr/local/etc/setpath.sh
#the next line enables the setup command
. /usr/local/etc/setups.sh
setup nt
print "FERMINT_DPOOL_DIR: $FERMINT_DPOOL_DIR\n"
needfile -stage -vsn YWF445 -tape 8mmda
let i=1
let numfiles=25
needfile -lb file_$i -vsn YWF445 -file $i -nw
while (($i <= $numfiles)) ; do
        if (($i < $numfiles))
        then
                let j=i+1
                needfile -lb file_$j -vsn YWF445 -file $j -nw
        fi
        needfile -wait file_$i
        print "Here we run my_program on $FERMINT_DPOOL_DIR/file_$i\n"
        if (($?))
        then
                print "We had a problem with file number $i\n"
        else
                print "We are done with file number $i\n"
        fi
        needfile -clear file_$i
        let i=i+1
done
exit
```

11.2 Hints on using needfile / fmss within FORTRAN program

```
   call getenv(name,whatcomesback)
   call system(anystringvaraiblewiththecommand)
```

12. Additional sources of related information

UNIX at Fermilab GU0001

FNALU User's Guide GU0008

The documentation listed below is available from: http://www-hppc.fnal.gov/mss

fmss User's Reference Manual

fmss Reference Page

needfile Reference Manual

mssout Reference Page

mssout Reference Manual